Communication

A Product Pose Tracking Paradigm Based on Deep Points Detection

Loukas Bampis * , Spyridon G. Mouroutsos and Antonios Gasteratos

School of Engineering, Democritus University of Thrace, 671 00 Xanthi, Greece; sgmour@ee.duth.gr (S.G.M.); agaster@pme.duth.gr (A.G.)
* Correspondence: lbampis@pme.duth.gr

Abstract: The paper at hand presents a novel and versatile method for tracking the pose of varying products during their manufacturing procedure. By using modern Deep Neural Network techniques based on Attention models, the most representative points to track an object can be automatically identified using its drawing. Then, during manufacturing, the body of the product is processed with Aluminum Oxide on those points, which is unobtrusive in the visible spectrum, but easily distinguishable from infrared cameras. Our proposal allows for the inclusion of Artificial Intelligence in Computer-Aided Manufacturing to assist the autonomous control of robotic handlers.

Keywords: deep learning; Computer-Aided Manufacturing; material processing; pose recognition

1. Introduction

During the last decade, significant effort has been given in facilitating contemporary digital technologies into the manufacturing procedure to comply with the Industry 4.0 scheme through Vertical Networking, Horizontal Integration, Through-Engineering, and early adaptation in Exponential Technologies [1,2]. Smart systems have been already introduced into the manufacturing procedure to increase the flexibility and productivity scale in different levels of infrastructure, such as remote control through the Internet of Things (IoT) [3], predictive maintenance [4], failure recovery from non-expert personnel [5,6], low-volume or high-variance production [7], workload scheduling [8], and more. With the means mentioned above, Computer-Aided Manufacturing (CAM) systems can be significantly improved to increase their autonomy and provide real-time analysis of their subjects using Artificial Intelligence (AI).

One of the most promising technologies that can be adopted in an Industry 4.0 ecosystem refers to the processing of visual data from low-cost camera sensors. During the last two decades, computer vision has reported tremendous achievements in automation and manufacturing, spanning from landmark/keypoint detection and extraction [9,10], to exploration [11] and novelty detection [12]. Combined with the cognitive capabilities of AI and Deep Learning (DL), such technologies enable smart systems to better interpret and interact with their environment. Thus, modern automation can perform complex tasks, while also handling unexpected events.

Pose estimation and tracking have been a challenging task that has provoked thought for many researchers. Accurately identifying the position and orientation of an object allows autonomous machines (e.g., robotic manipulators in a production line) to grasp and adequately handle it without compromising their integrity. Traditionally, the whole 3D figure of an item is estimated by using depth visual sensors, such as stereo or RGB-D cameras [13–15]. More recently, though, in order to provide cost-efficient solutions, the related literature has been focused on identifying an object's pose through single image instances, which can be acquired by monocular sensors [16]. This is achieved by either using Structure from Motion techniques via frames captured during different time

instances [17,18], or through DL methods [9,19–22], which can identify the full pose or representative local points of a given object.

In this paper, we present a solution for improving the operation of a smart assembly line, which requires the detection and tracking of products during manufacturing (e.g., [23]). Our proposal refers to an AI-enabled tool for generating representative tracking points to capture the position of different products and close the control loop of CAM systems. Instead of relying on fiducial [24] or other markers that affect an object's appearance, our conceptualization refers to the application of highly infrared (IR) reflective materials, such as Aluminum or Magnesium Oxide, to strategically selected by the AI [25] regions of each product. Such materials can be easily perceived by an IR camera sensor without significantly interfering with the final form of the product, facilitating the tracking and handling procedures of a modern production line with multiple robotic manipulators. Within the scope of this work, we consider the use of Aluminum Oxide, which constitutes a low-cost solution with high reflectivity on the IR spectrum [26]. Our approach realizes the identification of points within a pre-processing step without explicitly requiring the online deployment of the DL architecture during the manufacturing procedure. This allows the Aluminum Oxide markings to be planned beforehand and incorporated in the design procedure, as well as the use of direct methods for retrieving the pose of each object based on well-established 3D geometry techniques.

2. Proposed Approach

A DL network is proposed to automatically identify representative points to be processed with Aluminum Oxide (e.g., injections, paint, coating [27–30]) based on the architecture presented in [25]. An open source implementation of the used network can be found in the publicly available repository https://tinyurl.com/githubAttention, accessed on 25 May 2021. In this work, the DenseNet [31] model is used, according to which the extracted information from previous layers is accumulated in the following ones by propagating their respective feature maps. With the view to reduce the computational complexity of the original design, standard convolutional layers are substituted by depthwise separable ones [32]. Specifically, the model's stem consists of Dense Blocks with Inverted Residuals and Mish activation function [33]. Moreover, for sub-sampling feature maps, an anti-aliasing Blur Pooling filter is introduced, allowing the use of different kernel sizes [34]. The network's main building block, shown in Figure 1, consists of an Attention-Augmented Inverted Residual Block constructed around a standard residual one. Such a mechanism aggregates the similarity between profound query characteristics, and thus, multiple subspaces and spatial positions can be monitored by using a series of attention blocks. Finally, downsampling between Dense Blocks is achieved by a Transition Layer with pointwise convolutions to reduce the feature maps' depth, Blur Pooling, and batch normalization. The network's output corresponds to a set of k coordinates, each denoting a specific point on an object. Besides its competitive performance, this network is appealing for a dynamic pose detection system due to its single-stage end-to-end architecture and its capability to regress the representative points' coordinates within a pre-processing step; before the items reach the production stage.
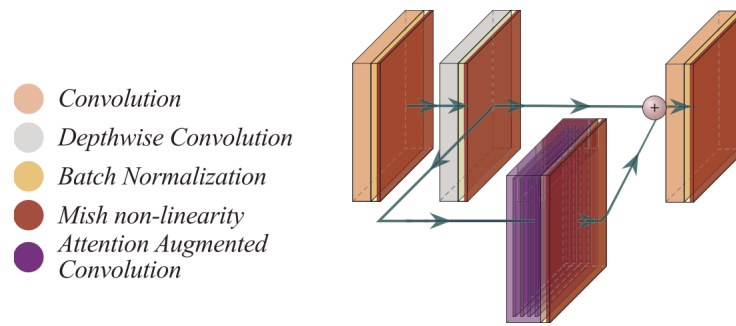
Figure 1. The structure of an Attention-Augmented Inverted Residual Block (with copyright permission from [25]).

Our approach refers to training the above network with a dataset of known manufacturing objects using their respective Computer-Aided Design (CAD) models. Consequently, the same network can automatically indicate representative tracking points for other products with different shapes. Our conceptualizing is based on the notion that the DL architecture has been trained to detect the most appropriate points from the learning objects, and it can transfer this knowledge to unknown items, as well. Given that CAD learning models can be transformed with respect to any desired viewing angle, the trained network will be able to identify the required amount of tracking points from multiple views. Thus, Aluminum Oxide can be placed accordingly to different objects' sides, allowing an IR camera to accurately identify their position.

More specifically, the proposed concept assumes a basic setup, such as the one presented in Figure 2, including a single or multiple IR camera sensors mounted on the ceiling, according to the production line's size that needs to be monitored. Using an IR source of lighting aimed at such a line, Aluminum Oxide points strongly reflect the light, and they are captured by the sensor. This allows for their detection and tracking on every frame of the respective video stream. Finally, the above 2D image points are associated with the known 3D object's mess (given from its CAD model), and the Perspective-n-Point (PnP) algorithm [35] is applied to estimate the relative 6-degrees-of-freedom transformation between the camera and each product. By assuming $n = 3$ for the PnP algorithm, four different solutions are obtained, which can be sorted out by a fourth point association. The P3P algorithm, applied on the simplistic setup presented in Figure 3 among the camera center P_0 and three 3D points P_i, $i \in \{1, 2, 3\}$, adheres to the following equation system:

$$\begin{cases} Y^2 + Z^2 - YZp - \mathrm{b}^2 = 0 \\ Z^2 + X^2 - XZq - \mathrm{c}^2 = 0, \\ X^2 + Y^2 - XYr - \mathrm{a}^2 = 0 \end{cases} \quad (1)$$

where:

$$\begin{cases} X = |P_0 - P_1|_2 \\ Y = |P_0 - P_2|_2, \\ Z = |P_0 - P_3|_2 \end{cases} \quad \begin{cases} p = 2\cos\angle P_2 P_0 P_3 \\ q = 2\cos\angle P_1 P_0 P_3, \\ r = 2\cos\angle P_1 P_0 P_2 \end{cases} \quad \begin{cases} \mathrm{a} = |P_1 - P_2|_2 \\ \mathrm{b} = |P_2 - P_3|_2. \\ \mathrm{c} = |P_1 - P_3|_2 \end{cases} \quad (2)$$

In the above, p, q, and r can be computed based on the 2D image point correspondences with the world's points P_i, while a, b, and c are known from the given CAD model. Furthermore, the notation $|...|_2$ denotes the L2-norm. By solving the equation system (1), we can obtain X, Y, and Z, which correspond to the depth information of the reference world points. Then, given the intrinsic camera parameters, the 3D points' coordinates are computed (P_i') with respect to the IR sensor's frame of reference. Finally, the relative rotation (R) and translation (T) between the camera and the product are recovered by solving:

$$P_i' = RP_i + T, \quad i \in \{1, 2, 3\}. \quad (3)$$

Considering that the IR sensor's position is known and fixed, each object's pose on the production line can be accurately retrieved and transformed on a global frame of reference. Please note that in the above procedure, one-to-one associations between the images' 2D and the products' 3D points need to be known. There are several approaches for obtaining this information (e.g., [9,36]); however, an efficient scheme is presented in [37], where a brute force searching strategy was constructed by inspecting every possible combination and permutation among 3 point detections. The resulting solutions of the P3P algorithm can then be sorted out based on the re-projection error from the rest of the detected 2D points. At first glance, such an approach may seem computationally intensive; however, one needs to consider that the number of image points is low ($k = 50$ for our case) and that each regressed coordinate set is associated with a specific point of an object determined in the training samples.
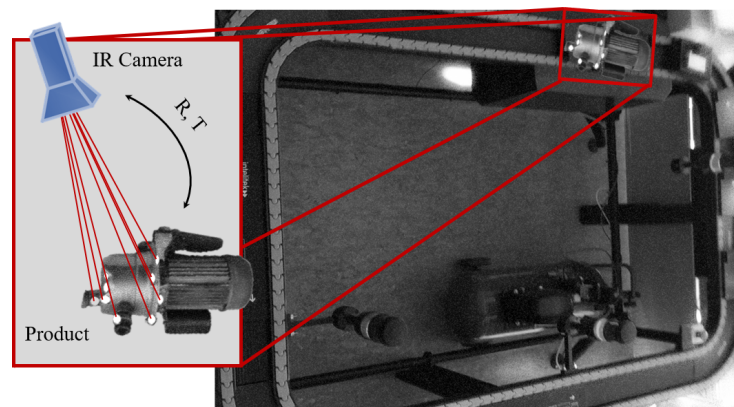

Figure 2. An indicative setup of the proposed concept for product tracking through an IR camera. The item is processed with an IR reflective material on points proposed by the Attention model [25]. Using the 3D-to-2D point associations, the relative rotation (R) and transformation (T) between the sensor and an item can be retrieved, allowing accurate pose estimation and effective manipulation in a production line.
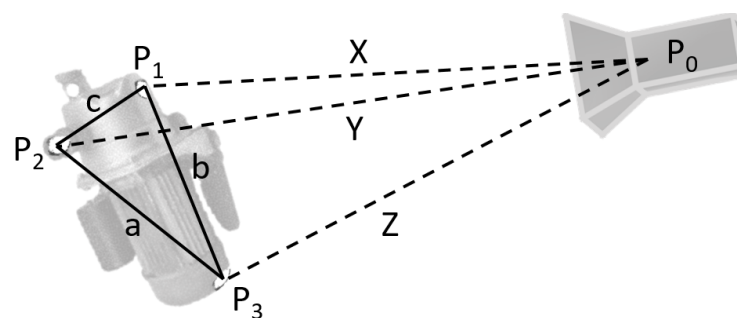

Figure 3. Schematic representation of the *P3P* problem.

3. Preliminary Results

A representative example of the obtained results is depicted in Figure 4. The network can automatically produce the heat-map of the products' most profound tracking points through the feature maps of the Attention-Augmented Inverted Residual Block. The presented outcome is obtained using the network provided in [25], which was originally trained on the PANOPTIC [38] dataset. More specifically, 165,000 learning subjects were used under the Stochastic Gradient Descent optimizer with a triangular policy of Cyclical Learning Rate [39]. The proposed DL architecture identifies regions of increased saliency on a novel object outside the training space, highlighting its versatility and generalization properties. With the view to further evaluate our proposed framework's performance, we conducted an additional round of experiments to measure the network's robustness for detecting salient points through repeatability. To that end, we applied a series of trans-

formations (including rotation, translation, and affine) on the item depicted in Figure 4, resulting in 300 different instances. Then, we deployed the Attention model on each of them and measured the detected points' repeatability using the Intersection over Union (IoU) metric. The obtained results are presented in Table 1 and indicate that most points can be effectively re-detected over multiple viewing angles. Please note that the network was set to detect the most prominent $k = 50$ points from each image.
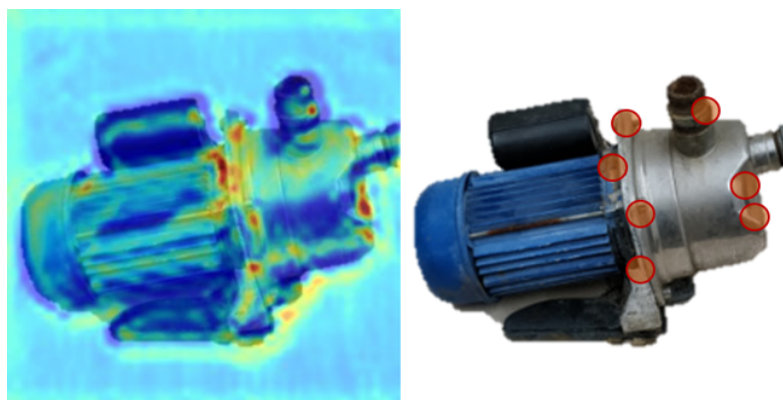

Figure 4. A representative example of applying our DL network on a novel object. Despite training the model on different items, the architecture can produce valid tracking points. (**Left**): Feature map of the Attention-Augmented Inverted Residual Block. (**Right**): Tracking points based on the produced output coordinates.

Table 1. Intersection over Union (IoU) for different number image transformations measuring the detected point's repeatability.

Number of transformed instances	50	100	150	200	250	300
IoU (%) among the detected points	91%	91%	87%	83%	80%	75%

It is worth noting that in real-world applications, the learning process needs to be implemented over CAD samples from the specific production line's targeted manufacturing products; however, such a learning sample is expected to be small. To that end, datasets such as the ABC [40] one can be used within a pre-training step to transfer-learn the general characteristics (e.g., geometry and appearance) of production items and restrain over-fitting.

Furthermore, the lightweight nature of the network, totaling only 1.9M mixed-precision parameters [41], offers a real-time performance of approximately 20 ms per frame when deployed on a Titan Xp GPU. In comparison, a hardware-accelerated version of hand-crafted keypoints, which do not incorporate any learned qualities for proper pose tracking, has been reported in [42] to be computed at 11 ms per input sample. This real-time performance allows for the detection of points to be refined with Aluminum Oxide even during the manufacturing procedure; however, one can assign this process to earlier stages, such as the products' design.

4. Conclusions

Our approach proposes the application of a state-of-the-art deep network on the manufacturing procedures. Realizing the identification of points for product pose tracking as a highly dynamical system, we evaluate our technique in terms of repeatability, establishing its robustness against different viewing angles. This is due to the rich information contained within the DL model's multi-layer structure, its ability to describe highly correlated input-output variables, and the use of high-dimensional learning data. Such a framework paves the way to increase a manufacturing facility's level of automation, leading to more effective control in modern industrial environments and remote inspection of the production

line within an IoT paradigm. As part of our future, we plan to evaluate a complete system based on our proposal to showcase fully solidified results in an industrial environment.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kagermann, H.; Helbig, J.; Hellinger, A.; Wahlster, W. *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry; Final Report of the Industrie 4.0 Working Group*; Forschungsunion: Munich, Germany 2013.
2. Schläpfer, R.C.; Koch, M.; Merkofer, P. *Industry 4.0 Challenges and Solutions for the Digital Transformation and Use of Exponential Technologies*; Finance, Audit Tax Consulting Corporate: Zurich, Switzerland, 2015.
3. Xu, L.D.; Xu, E.L.; Li, L. Industry 4.0: State of the Art and Future Trends. *Int. J. Prod. Res.* **2018**, *56*, 2941–2962. [CrossRef]
4. Wang, K. Intelligent Predictive Maintenance (IPdM) System-Industry 4.0 Scenario. *WIT Trans. Eng. Sci.* **2016**, *113*, 259–268.
5. Subakti, H.; Jiang, J.R. Indoor Augmented Reality Using Deep Learning for Industry 4.0 Smart Factories. In Proceedings of the IEEE Annual Computer Software and Applications Conference, Tokyo, Japan, 23–27 July 2018; Volume 2, pp. 63–68.
6. Konstantinidis, F.K.; Kansizoglou, I.; Santavas, N.; Mouroutsos, S.G.; Gasteratos, A. MARMA: A Mobile Augmented Reality Maintenance Assistant for Fast-Track Repair Procedures in the Context of Industry 4.0. *Machines* **2020**, *8*, 88. [CrossRef]
7. Olsen, T.L.; Tomlin, B. Industry 4.0: Opportunities and Challenges for Operations Management. *Manuf. Serv. Oper. Manag.* **2020**, *22*, 113–122. [CrossRef]
8. Zhang, J.; Ding, G.; Zou, Y.; Qin, S.; Fu, J. Review of Job Shop Scheduling Research and its New Perspectives Under Industry 4.0. *J. Intell. Manuf.* **2019**, *30*, 1809–1830. [CrossRef]
9. Lambrecht, J. Robust Few-Shot Pose Estimation of Articulated Robots Using Monocular Cameras and Deep-Learning-Based Keypoint Detection. In Proceedings of the IEEE International Conference on Robot Intelligence Technology and Applications, Daejeon, Korea, 1–3 November 2019; pp. 136–141.
10. Nilwong, S.; Hossain, D.; Kaneko, S.i.; Capi, G. Deep Learning-Based Landmark Detection for Mobile Robot Outdoor Localization. *Machines* **2019**, *7*, 25. [CrossRef]
11. Mur-Artal, R.; Tardós, J.D. ORB-SLAM2: An Open-Source SLMA System for Monocular, Stereo, and RGB-D cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
12. Bampis, L.; Gasteratos, A.; Boukas, E. CNN-based novelty detection for terrestrial and extra-terrestrial autonomous exploration. *IET Cyber-Syst. Robot.* **2021**. [CrossRef]
13. Cosenza, C.; Nicolella, A.; Esposito, D.; Niola, V.; Savino, S. Mechanical System Control by RGB-D Device. *Machines* **2021**. [CrossRef]
14. Xu, H.; Chen, G.; Wang, Z.; Sun, L.; Su, F. RGB-D-Based Pose Estimation of Workpieces with Semantic Segmentation and Point Cloud Registration. *Sensors* **2019**, *19*, 1873. [CrossRef] [PubMed]
15. Lima, J.P.; Roberto, R.; Simões, F.; Almeida, M.; Figueiredo, L.; Teixeira, J.M.; Teichrieb, V. Markerless Tracking System for Augmented Reality in the Automotive Industry. *Expert Syst. Appl.* **2017**, *82*, 100–114. [CrossRef]
16. Pavlakos, G.; Zhou, X.; Chan, A.; Derpanis, K.G.; Daniilidis, K. 6-DOF Object Pose from Semantic Keypoints. In Proceedings of the IEEE International Conference on Robotics and Automation, Marina Bay Sands, Singapore, 29 May–3 June 2017; pp. 2011–2018.
17. Bampis, L.; Gasteratos, A. Revisiting the Bag-of-Visual-Words Model: A Hierarchical Localization Architecture for Mobile Systems. *Robot. Auton. Syst.* **2019**, *113*, 104–119. [CrossRef]
18. Ni, T.; Shi, Y.; Sun, A.; Ju, B. Simultaneous Identification of Points and Circles: Structure from Motion System in Industry Scenes. *Pattern Anal. Appl.* **2021**, *24*, 333–342. [CrossRef]
19. Haochen, L.; Bin, Z.; Xiaoyong, S.; Yongting, Z. CNN-Based Model for Pose Detection of Industrial PCB. In Proceedings of the IEEE International Conference on Intelligent Computation Technology and Automation, Changsha, China, 9–10 October 2017; pp. 390–393.
20. Brucker, M.; Durner, M.; Márton, Z.C.; Balint-Benczedi, F.; Sundermeyer, M.; Triebel, R. 6DoF Pose Estimation for Industrial Manipulation Based on Synthetic Data. In *International Symposium on Experimental Robotics*; Springer: Cham, Switzerland, 2018; pp. 675–684.
21. Patel, N.; Mukherjee, S.; Ying, L. EREL-Net: A Remedy for Industrial Bottle Defect Detection. In *International Conference on Smart Multimedia*; Springer: Cham, Switzerland, 2018; pp. 448–456.
22. Chen, X.; Chen, Y.; You, B.; Xie, J.; Najjaran, H. Detecting 6D Poses of Target Objects From Cluttered Scenes by Learning to Align the Point Cloud Patches With the CAD Models. *IEEE Access* **2020**, *8*, 210640–210650. [CrossRef]

23. Konstantindis, F.K.; Gasteratos, A.; Mouroutsos, S.G. Vision-Based Product Tracking Method for Cyber-Physical Production Systems in Industry 4.0. In Proceedings of the IEEE International Conference on Imaging Systems and Techniques, Krakow, Poland, 16–18 October 2018; pp. 1–6.

24. Botta, A.; Quaglia, G. Performance Analysis of Low-Cost Tracking System for Mobile Robots. *Machines* **2020**, *8*, 29. [CrossRef]

25. Santavas, N.; Kansizoglou, I.; Bampis, L.; Karakasis, E.; Gasteratos, A. Attention! A Lightweight 2D Hand Pose Estimation Approach. *IEEE Sens. J.* **2020**. [CrossRef]

26. Melnichuk, O.; Melnichuk, L.; Venger, Y.; Torchynska, T.; Korsunska, N.; Khomenkova, L. Effect of Plasmon–Phonon Interaction on the Infrared Reflection Spectra of $Mg_xZn_{1-x}O/Al_2O_3$ Structures. *J. Mater. Sci. Mater. Electron.* **2020**, *31*, 7539–7546. [CrossRef]

27. Melcher, R.; Martins, S.; Travitzky, N.; Greil, P. Fabrication of Al2O3-Based Composites by Indirect 3D-Printing. *Mater. Lett.* **2006**, *60*, 572–575. [CrossRef]

28. Huang, S.; Ye, C.; Zhao, H.; Fan, Z. Additive Manufacturing of Thin Alumina Ceramic Cores Using Binder-Jetting. *Addit. Manuf.* **2019**, *29*, 100802. [CrossRef]

29. Lorenzo-Martin, C.; Ajayi, O.; Hartman, K.; Bhattacharya, S.; Yacout, A. Effect of Al_2O_3 Coating on Fretting Wear Performance of Zr Alloy. *Wear* **2019**, *426*, 219–227. [CrossRef]

30. Misra, V.C.; Chakravarthy, Y.; Khare, N.; Singh, K.; Ghorui, S. Strongly Adherent Al_2O_3 Coating on SS 316L: Optimization of Plasma Spray Parameters and Investigation of Unique Wear Resistance Behaviour Under Air and Nitrogen Environment. *Ceram. Int.* **2020**, *46*, 8658–8668. [CrossRef]

31. Huang, G.; Liu, Z.; Pleiss, G.; Van Der Maaten, L.; Weinberger, K. Convolutional Networks with Dense Connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef] [PubMed]

32. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.

33. Misra, D. Mish: A Self Regularized Non-Monotonic Activation Function. *arXiv* **2019**, arXiv:1908.08681.

34. Zhang, R. Making Convolutional Networks Shift-Invariant Again. In Proceedings of the International Conference on Machine Learning, Los Angeles, CA, USA, 9–15 June 2019; pp. 7324–7334.

35. Gao, X.S.; Hou, X.R.; Tang, J.; Cheng, H.F. Complete Solution Classification for the Perspective-Three-Point Problem. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 930–943.

36. Kubíčková, H.; Jedlička, K.; Fiala, R.; Beran, D. Indoor Positioning Using PnP Problem on Mobile Phone Images. *Int. J. Geo-Inf.* **2020**, *9*, 368. [CrossRef]

37. Faessler, M.; Mueggler, E.; Schwabe, K.; Scaramuzza, D. A Monocular Pose Estimation System Based on Infrared LEDs. In Proceedings of the IEEE International Conference on Robotics and Automation, Hong Kong, China, 31 May–7 June 2014; pp. 907–913.

38. Simon, T.; Joo, H.; Matthews, I.; Sheikh, Y. Hand Keypoint Detection in Single Images Using Multiview Bootstrapping. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1145–1153.

39. Smith, L.N. Cyclical Learning Rates for Training Neural Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Santa Rosa, CA, USA, 24–31 March 2017; pp. 464–472.

40. Koch, S.; Matveev, A.; Jiang, Z.; Williams, F.; Artemov, A.; Burnaev, E.; Alexa, M.; Zorin, D.; Panozzo, D. ABC: A Big CAD Model Dataset for Geometric Deep Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9601–9611.

41. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. Mixed Precision Training. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.

42. Bampis, L.; Amanatiadis, A.; Gasteratos, A. Fast Loop-Closure Detection Using Visual-Word-Vectors from Image Sequences. *Int. J. Robot. Res.* **2018**, *37*, 62–82. [CrossRef]